VA Capitol Health Care Network VISN 5

10/10/2002

Introduction to the IRIS Database for MIRECC Investigators

I. Background on IRIS

The integrated research information system (IRIS) is a comprehensive system that combines tailor-made software (Oracle platform) for managing research protocols (e.g. enrollment, scheduling, and tracking) with a system for efficient collection, capture, storage, and retrieval of data. We refer to the entire system simply as "IRIS." IRIS was developed and is maintained by the Informatics Group, Department of Epidemiology to meet the needs of a broad range of researchers and research centers at, or associated with, UMAB and VAMCS Baltimore. There are two types of data used by the system: (1) data for managing the recruitment, scheduling, and monitoring of subjects participating in a study, and (2) data on subjects used to answer the scientific questions specified in the research protocol. This document discusses the latter.

We can maximize the benefits and capabilities of IRIS by structuring and standardizing MIRECC research protocol and data collection procedures to integrate with the system. The goals and policies of the MIRECC Research Core and of the individual studies should both be taken into consideration. In the initial year of development for the MIRECC's use of IRIS, many meetings have been held and decisions made regarding data standards and practices.

As an example, one MIRECC goal is to have the capability to construct data sets that combine common measures on subjects from multiple studies. To achieve this goal, data standards (such as using the same missing codes) have been put in place to assure consistency. (Follow this link to learn about IRIS data standards.) At the instrument level, another form of standardization toward this goal is to encourage investigators to use the same assessments, so that data can be pooled from several studies.

To prepare a study for data collection, read "Project Preparation" under "Information for Researchers" on the VISN 5 MIRECC Assessment Core Homepage.

2. Tutorial on the IRIS database (containing data on research subjects)

The IRIS database is made up of many tables linked together in what is known as a ***relational*** database. Each assessment or teleformed collection of questions and information has its own table. Each table is structured like an Excel or SPSS spreadsheet. Rows are called ***records***; columns are ***variables***. There is, for example, a PANSS table in the database. All PANSS data (captured from the PANSS teleform) across MIRECC studies are stored in the same table. MIRECC staff and data analysts, however, will retrieve data by study (if they have authorization). In most tables, each record contains data on an individual subject. If repeat assessments are conducted, the table will contain a separate record for each date. In special cases, tables may be organized slightly different, for example, when multiple raters rate the same subject.

Usually, only raw item data are stored, not total scores. Total scores would be computed after retrieval of the table using SAS or SPSS (with link-up to the database).

An exception to this are neuropsychological score totals computed on stand-alone systems.

There are two types of variables in each table – ***key*** variables and ordinary variables. Key variables are used to link data across tables. The most important key variable is the participant id number (PID). The four key variables used in the IRIS database (in hierarchical order) are: RPN number (study id), site id number, PID, and assessment date. The corresponding teleform fields for these four variables appear at the top of the first page of each teleform. The first four variables of each table also begin with the 4 key variables. Data in different tables are therefore collected, linked, and organized by these key variables. The site id number allows for multi-site studies. Different clinics, offices, meeting rooms on the UMAB campus are ***not*** considered different sites.

An arbitrary PID number for a subject is normally computer generated at the time the subject is enrolled through a web-based enrollment screen. There is an option to type in pre-assigned PID's should this be necessary, for example for multi-site studies. PID's are study specific so that a subject participating in, say, two studies will be assigned two distinct PID's. The same PID can be used in multiple studies. Behind the scenes, IRIS also assigns participants a "unique ID" (UID), which follows the subject across studies. This ID is not accessible to users of the system, but will be used by the system, for example, if there is a need to combine data on subjects who have participated in more than one study (given IRB approval).

Teleforms are the teleformed versions of assessment instruments or other collections of data under a specific category. For example the MIRECC demographic form collects data on psychiatric history and demographic characteristics. Teleforms are made up of items, each of which corresponds to a separate piece of data. There are generally two types of items on the teleforms: darken-the-bubble multiple choice and write-in boxes. There is a one-to-one correspondence between each item and a variable (column) or ***data field*** in the corresponding like-named data table.

Forms cannot be scanned unless the key variable data fields (at the top of the first page of each form) are completed. For convenience and efficiency, a study can pre-print their RPN number on each of their teleforms so that this unchanging code doesn't have to repeatedly be filled-in. It is the responsibility of the principal investigator to keep tight control of pre-printed forms. Otherwise, another study might inadvertently use the forms and data would be incorrectly captured under the wrong study.

In constructing new teleforms, in some cases it's important to consider the use of the form in other studies to the extent possible. Constructing differing versions of a form with slight differences is inefficient, expensive, and strongly discouraged. Go to the Assessment Core Homepage to view (or print) any currently available teleformed instrument.

A new table is constructed in the IRIS database when a new teleform is developed. Specifications for each data field are set by the ***data dictionary***. The data dictionary is used by the system to translate user-friendly darken-the-bubble choices to coded data stored in the database. Characteristics specified in the data dictionary include variable name, description, and allowable range. In the scanning process, out-of-range values are flagged forcing the person scanning to resolve the discrepancy. The data dictionary is determined by investigators and the assessment core director by filling out a data

dictionary form (see link in “Project Preparation” document on the VISN 5 MIRECC Assessment Core Homepage.)

There are other types of tables in the IRIS database besides those constructed during the teleforming process. The “encounter” table corresponding to the encounter form is one example. The encounter form and table track which dates correspond to which assessment points or encounter numbers (per the protocol). It also allows the recording of one encounter spread over multiple dates and to assign assessment data from one encounter in one study to another encounter in a different study, if the subject is participating in more than one study and the same assessment is being used. All studies are currently required to use the encounter form (see “Data Standards for Investigators”)